Assessing Fair Policing in Austin, TX

Team FunkyStats

4/18/2021

Introduction

This paper investigates racial disparities in traffic stops by the Austin Police Department. Using data available from Austin Open Data, a Texas government-run data portal, and from the Stanford Open Policing Project, we evaluate these disparities using models derived from the "hit rate" and the effect of the "veil of darkness," two often-cited methods for assessing fair policing. Our main report consists of three parts. First, we conduct an exploratory data analysis to get a big picture of policing in Austin. Second, we use various modeling strategies to assess the severity of racial disparities. Third, we propose a measure of fairness based on the differences in the posterior median hit rate among individual police officers.

Available Data

The primary data set is from the Stanford Open Policing Project[1] (from here on referred to as the Stanford data). This data set record stops made by the APD a roughly ten year period (2006.01.01 - 2016.06.30) and contains information such as the date of the stops, the subject's race, whether the person was searched or frisked, whether any contraband were found. Notably, this data lacks information about the time or place of the stops. Because the 2016 data is incomplete, we focus on the data for which we have complete years (2006-2015) for the bulk of the analysis; this contains 463,944 stops.

[1] Stanford Open Policing Project (OPP): https://openpolicing.stanford.edu/data/

Our secondary data set is from the 2019 Racial Profiling report, available from Austin Open Data (and hereafter referred to as the RP data). This data set contains similar information as the Stanford data, with additional information about the event time, location, and officer race. Notably, the race of the subject is missing from this data.

Lastly, we use US census demographic data. Specifically, we use 2017 5-year American Community Survey zip-code-level data with the Stanford data, and 2019 census population data for 2019 Austin RP data. In addition, we also refer to the racial profiling reports from the Austin Police Department.

Summary Statistics

Summary statistics for the Stanford data are as follows.The statistics reported cover all available data (2006.01.01 - 2016.06.30). Unique officer IDs are available but not shown here.
% latex table generated in R 4.0.5 by xtable 1.8-4 package % Mon Apr 19 13:16:56 2021

	nobs	nmis	uniq	mean	SD	min	25

% latex table generated in R 4.0.5 by xtable 1.8-4 package % Mon Apr 19 13:17:05 2021

	nobs	nmis	uniq	mean	SD	min	25%
contraband_found	19256	0	2	0.25	0.43	0.00	0.00
contraband_drugs	19256	0	2	0.01	0.12	0.00	0.00
contraband_weapons	19256	0	2	0.05	0.21	0.00	0.00
frisk_performed	19256	0	2	0.51	0.50	0.00	0.00

Table 1: Subject race.

Race	n	percent
asian/pacific islander	13167	0.0272466
black	72324	0.1496607
hispanic	123943	0.2564764
other	2626	0.0054340
unknown	3135	0.0064873
white	268058	0.5546950
NA	2	NA

Table 2: Search basis.

Search basis	n	percent
consent	3195	0.1659223
other	276	0.0143332
plain view	152	0.0078936
probable cause	15633	0.8118509
NA	463999	NA

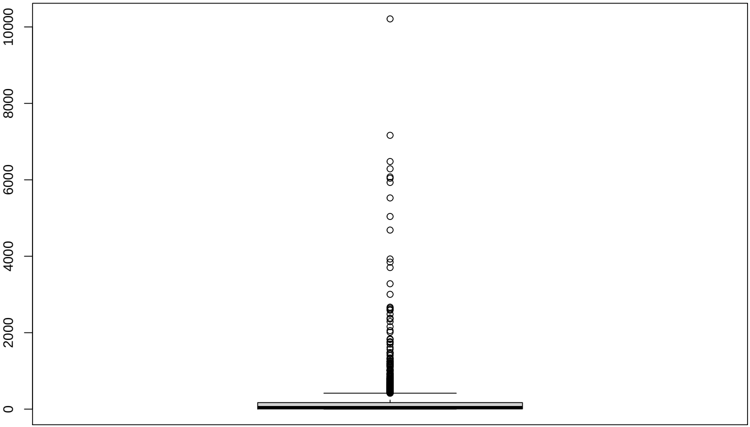


Figure 1: Distribution of stops by unique officer ID

We note that the distribution of stops per officer has an extremely

Exploratory Analysis

We first examine the count of stops by race during 2006-2015 (Table 3), using the Stanford Data. It is notable that over half of the stops involved were of white subjects, about four times the number of stops of Black people. According to 5-year census data, the white population in Austin (445,269) is almost 7 times than the black population (66,724) — a classic Simpson's paradox. Examining figure 2, we can see that at least for Black, Hispanic and white drivers, the annual trends are very different by race.

Driver Race	Counts	Proportion
asian/pacific	11658	0.033
black	52381	0.147
hispanic	765707	0.215
other	2105	0.006
unknown	2622	0.007
white	211588	0.593

Table 3: Proportion of stops by race during 2006-2015

Because neither of our data sets contain both driver races and stop time information, we venture to an indirect way by measuring the racial population in different areas through zip codes by 2019 RP data provided by Austin Police Department. In order to accurately distinguish the daytime and nighttime, we compute the daily subset and dusk time for Austin in 2019. In Table 5 we can see earliest sunset in 2019 was at around 17:32 in early December and it goes fully dark in 26 minutes. The latest sunset time was around 20:38 late June and it was fully dark after 28 minutes.

Date	Sunset	Dusk	Sunset Minute	Dusk Minute
2019-12-02	17:31:42	17:57:48	1051	1077
2019-12-01	17:31:45	17:57:48	1051	1077
2019-06-30	20:37:58	21:05:27	1237	1265
2019-06-29	20:37:56	21:05:27	1237	1265

Table 5: Minimum and maximum dusk time during the 2019 in Austin

	Day	Night
BDA	124	126
WDA	2037	2216

We denote the stops happening before sunset as Daytime Stop, and the stop happening after the dusk as Nighttime Stop. We do not consider the stops happening between the sunset and dusk in this study. According to ZIP codes and the corresponding demographic data, we consider the areas that have more black people as black dominant area (BDA), and the areas consist of more white people as white dominated area (WDA). For simplicity of the analysis, here we consider only the black and the white population groups. Hence, each zip code is regarded as a location with label as white (WDA) or black (BDA).

We record the stops happening in each zip codes into two categories: daytime stops or nighttime stops, and we treat two rows in Table 6 as independent binomial samples. Of $n_1 = 250$ recorded stops in black dominated area, 124 stops happened during the daytime, a proportion of $p_1 = 124/250 = 0.496$. Of $n_2 = 5153$ recorded stops in white dominated area, 124 stops happened during the daytime, a proportion of $p_2 = 2937/5153 = 0.570$. The sample difference of proportions is 0.074. We obtain Fisher's exact test for testing null hypothesis of independence of the two rows with p value of 0.02,

Modeling

Logistic Regression

Logistic Regression for Frisk Rate

Our descriptive analysis shows that black people in Austin seem to be more likely to be stopped by the police. We want to answer the question, given a person is stopped, what factors may impact the likelihood of that person being frisked? To investigate this, we fit a logistic regression model with `frisk` as the dependent variable and `race`, `age`, and `sex`.

$$\text{Logit}[\text{P(Being Frisked)}] = \beta_0 + \beta_1\text{Race} + \beta_2\text{Age} + \beta_3\text{Sex}$$

Results can be found below.

Logistic Regression for Contraband found

We want to investigate how likely contraband items are found when searching is performed. This is equivalent to calculating hit rate defined in section 2.2.2. We argue that if racial bias does not exist, the hit rate should be equal for all races. In other words, we expect

Other Results

Contraband found	Drugs	Weapons	Others
(Intercept)	$-5.24^{***}(1.00)$	$-3.61^{***}(0.45)$	$-2.38^{***}(0.26)$
Black	$1.10(1.01)$	$0.32(0.46)$	$0.99^{***}(0.26)$
Hispanic	$1.21(1.01)$	$0.36(0.46)$	$1.04^{***}(0.26)$
White	$0.73(1.01)$	$0.98^{*}(0.46)$	$0.72^{**}(0.26)$
Other	$0.97(1.42)$	$0.47(0.74)$	$0.87^{*}(0.40)$
Unknown	$-11.33(280.85)$	$0.04(0.85)$	$-0.23(0.53)$

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Logistic model for frisk rate vs. race, age, and sex

term	estimate	std.error	statistic	p.value
(Intercept)	-2.984	0.102	-29.402	0.000
subject_raceblack	1.503	0.100	15.044	0.000
subject_racehispanic	1.310	0.099	13.214	0.000
subject_racewhite	0.719	0.099	7.260	0.000
subject_raceother	0.617	0.180	3.434	0.001
subject_raceunknown	0.647	0.183	3.544	0.000
subject_age	-0.046	0.001	-51.125	0.000
subject_sexfemale	-1.643	0.036	-45.311	0.000

Table 8: Logistic model for contraband found vs. race

term	estimate	std.error	statistic	p.value
(Intercept)	-1.988	0.222	-8.944	0.000
subject_raceblack	0.899	0.225	3.999	0.000
subject_racehispanic	0.941	0.224	4.202	0.000
subject_racewhite	0.826	0.224	3.686	0.000

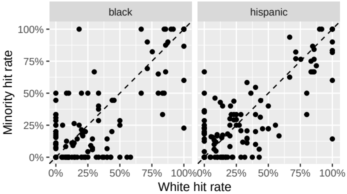


Figure 3: Hit rates for individual officers.

The above plot shows the hit rates for individual officers. An officer with an identical hit rate for white and minority subpopulations would be on the 45-degree line. Visually, it is difficult to determine a systematic trend, although it is clear that particular officers have hit rates that differ substantially by subpopulaton. It should be noted that the hit rate is highly variable with small sample sizes.

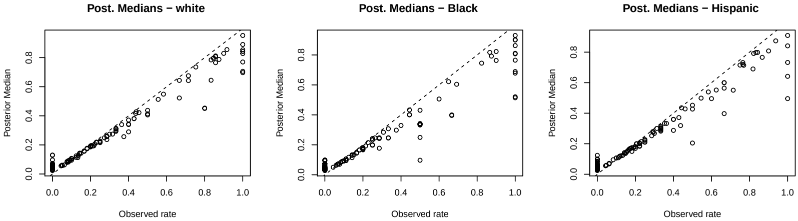
We proceed using a Bayesian hierarchical model. Under this model, we treat individual officers as belonging to a population of players and we seek to model both the hit rates of the officers and the variation of this population. This permits *partial pooling*, by which individual hit rates are biased towards the population average by an amount determined by the estimate of the population. For each officer, we consider three hit rates, one each for white, Black, and Hispanic subpopulations. We accomplish this by fitting separate logistic mixed effects models for each race, each with a weakly informative Normal prior on the log-odds with mean -1.2 and standard deviation 1.

Specifically, let θ_{jr} be the hit rate for for officer j and race r, y_{jr} be the number of hits, and K_{jr} the number of frisks. In the following, because we fit separate models, we assume for example $r = black$ and drop the r subscript. Assuming each officer's searches are independent Bernoulli trials

$$p(y_j|\theta_j) = \mathrm{Binomial}(y_j|K_j, \theta_j)$$

We reparametrize the model in terms of the log-odds, α:

The the following, the effects of partial pooling are evident: the posterior medians are baised towards the population average. Practically, this means that observed hit rates equal to zero have posterior medians that are small but positive, and perfect (or near-perfect) observed hit rates have somewhat smaller posterior medians.



Because part of this project is to "operationalize" fairness, we devised a measure by which the above posteriors can be converted into a rough "fairness score." Because an officer that uses the same evidence threshold when deciding whether to frisk a subject regardless of race should have roughly equal hit rates for all three subpopulations, we reason that such an officer should have posterior medians that are close to each other for the three subpopulations. So, one can calculate a simple sum of squares statistic for each officer. Specifically, letting m_{jr} be the posterior median for officer j and race r, the sum of squares statistic S_j is

$$S_j = \sum_r (m_{jr} - \bar{m}_j)^2$$

where $\bar{m}_j$ is the average of the three medians. Of course, this measure disregards all other information that could be gleaned from the posterior; an alternative might calculate the overlap between the posterior densities. However, we think this measure is relatively easy to understand and implement.

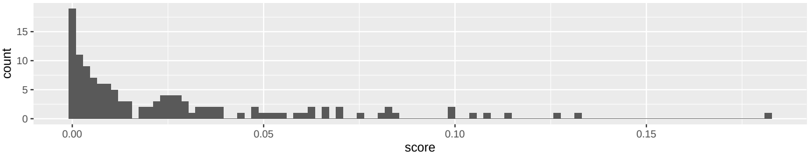


Figure 4: Fairness scores for the officers under consideration. Lower scores indicate hit rates are more similar.

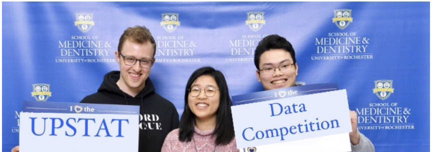
Table 10: Highest 5 scores.

ID	obs.w	obs.b	obs.h	post.w	post.b	post.h	count.
1504c3bc16	1.000	0.227	0.143	0.697	0.215	0.142	
3392a495a3	0.400	0.000	0.545	0.371	0.046	0.499	1
50f70c6ecb	0.583	0.000	0.167	0.549	0.064	0.162	1
bab7c2acaf	0.833	0.333	0.750	0.644	0.247	0.715	
dd9c1003d5	0.556	0.000	0.250	0.513	0.043	0.210	

Conclusion

In this study, we evaluate the fairness of traffic stops during the past two decades through three tests, namely benchmark test, outcome test and veil of darkness test. We found that the racial disparity in policing exists and is present in different scales spatially and temporarily. We also explore the causal confounding issues through logistic regression, finding that black and Hispanic people are more likely to be frisked and found with contraband items that are neither drugs or weapons.

Through the investigation of the hit rate via Bayesian hierarchical modeling, we obtained posteriors for the hit rate for each officer in a subset of the data. Using the medians for these posteriors, we devised a "fairness score," a tool we believe could be used to identify officers with racially disparate patterns of traffic stops.



References

Grogger, Jeffrey, and Greg Ridgeway. 2006. *Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness*. Santa Monica, CA: American Statistical Association.

"Hierarchical Partial Pooling for Repeated Binary Trials." n.d. https://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736–45. https://doi.org/10.1038/s41562-020-0858-1.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. 2017. "THE Problem of Infra-Marginality in Outcome Tests for Discrimination." *The Annals of Applied Statistics* 11 (3): 1193–1216. http://www.jstor.org/stable/26362224.